Clinical Research Databases

Stephen Johnson, PhD
Associate Professor
Medical Informatics
Columbia University

Presentation to NIH, August 30, 2000

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Clinical Databases

- Ancillary: access by specimen, image or prescription (transactional)
- Patient Care: access by single patient for current encounter or longitudinal view of care (historical)
- Research: aggregation over groups of subjects (analytical)
- Administrative: resource utilization and cost (analytical)

Clinical Center Goals

- Collect data once for patient care and for research
- Collect data using protocols
- Facilitate access to data by researchers
- Monitor patient safety

Clinical Center Goals

- Collect managerial and financial data
- Integrate clinical research outputs with resource allocation inputs
- Inform decision-making processes
- Measure performance and cost
- Integrate data between and among departments

Institutions

- Massachusetts General – COSTAR
- Duke University – Perinatal Repository
- Regenstrief Institute – Medical Record System
- Pittsburgh – MedisGroups
- Yale – ACT Database
- University of Virginia – Clinical Data Repository

Columbia Databases

- Patient Care: Clinical Repository
- Research: Clinical Data Warehouse
- Integration of administrative data (charges)
- 100 Gigabytes
- 10 years of data
- Open Architecture (multiple vendors)
- Informatics support

State of the Art

- Few institutions
- Almost no publications
- Little vendor support

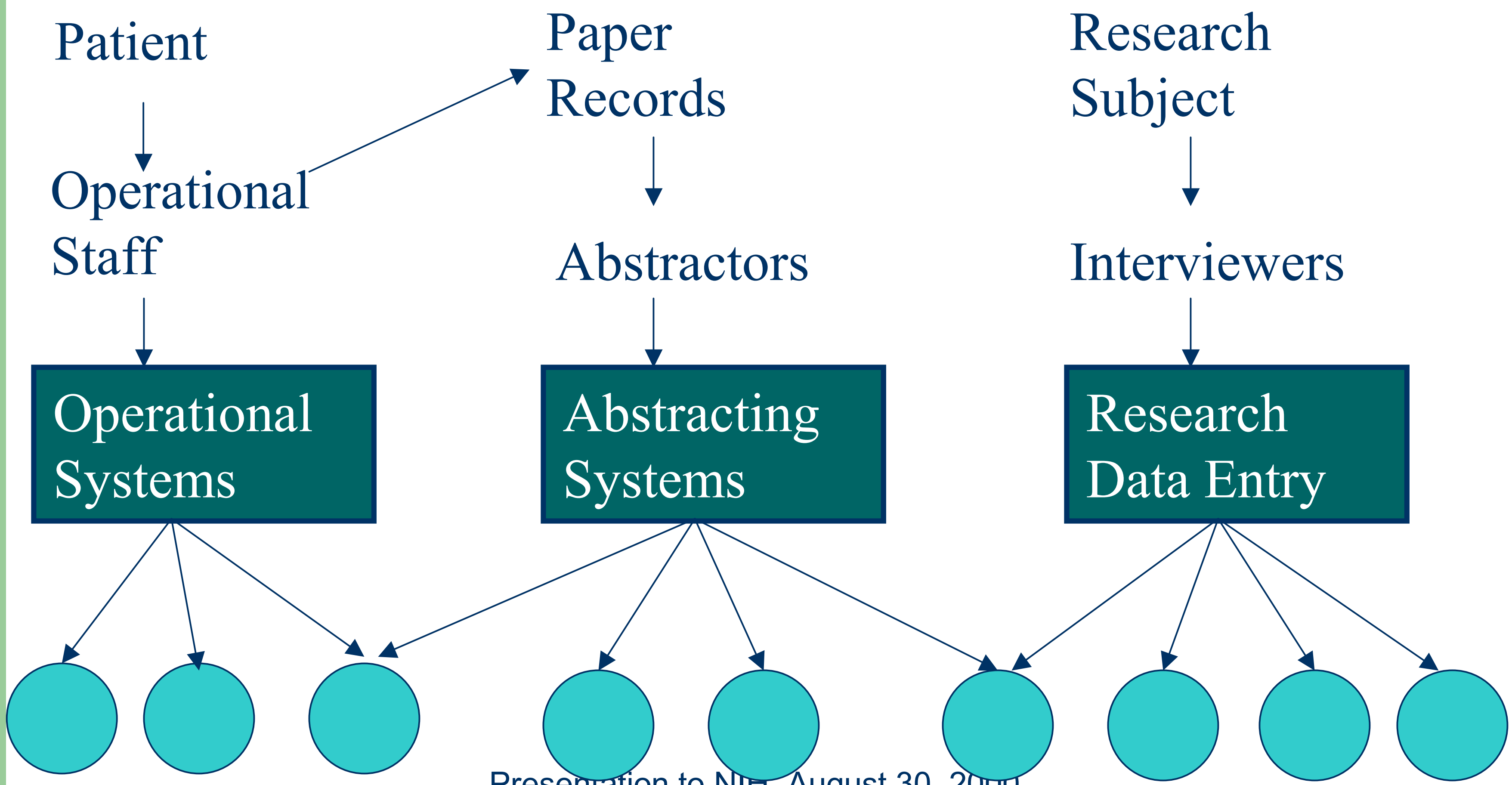
Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Barriers to Research Databases

- Patient care: shared medical records
- Research: study-specific records
- Special needs of research
- Ownership of data
- Lack of automation in health care

Proliferation of Databases



Patient

Operational Staff

Paper Records

Abstractors

Research Subject

Interviewers

Operational Systems

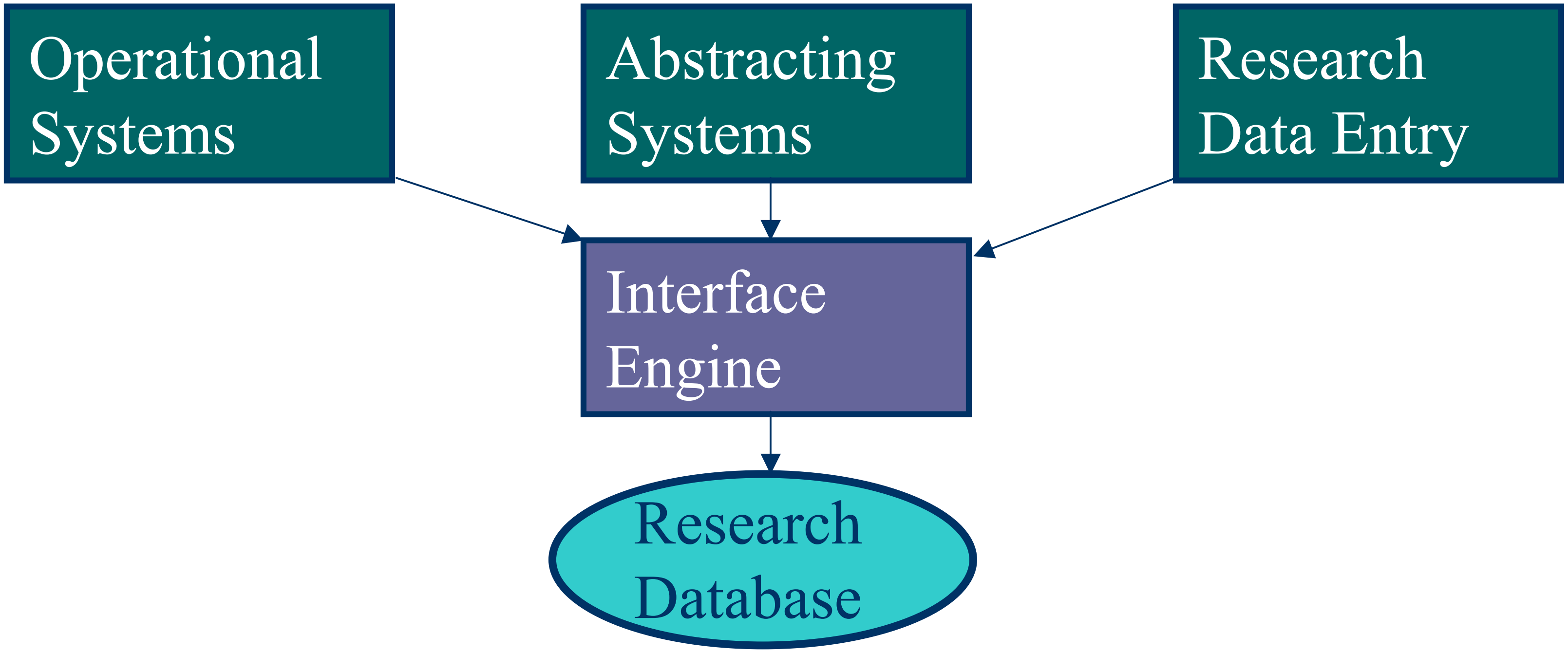
Abstracting Systems

Research Data Entry

Effects of Proliferation

- Redundant data collection
- Lack of integration
- Incompatible data
- Failure to leverage valuable data

Research Database Architecture



Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

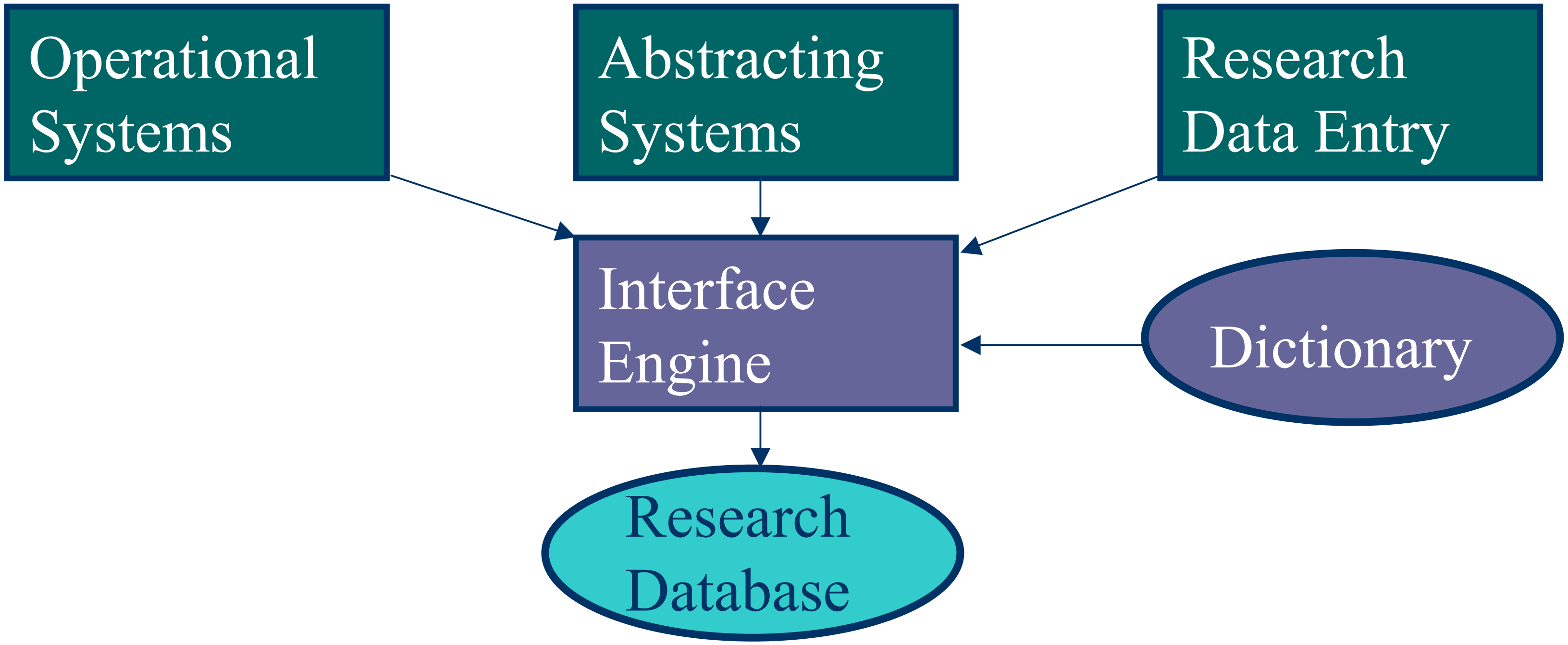
Clinical Research Variables

- Demographics – age, sex, residence
- Risk factor – behavior, environmental exposure
- Diagnosis – diseases, problems
- Finding – symptom, test result
- Treatment – medication, surgery
- Health Status – mortality, functional status
- Cost – charges, resource utilization, personnel

Bias in Variables

- Lack of precise definitions
- Differing granularities of interest
- Derivation from other data
- Conditions of collection

Data Dictionary



Operational Systems

Abstracting Systems

Research Data Entry

Interface Engine

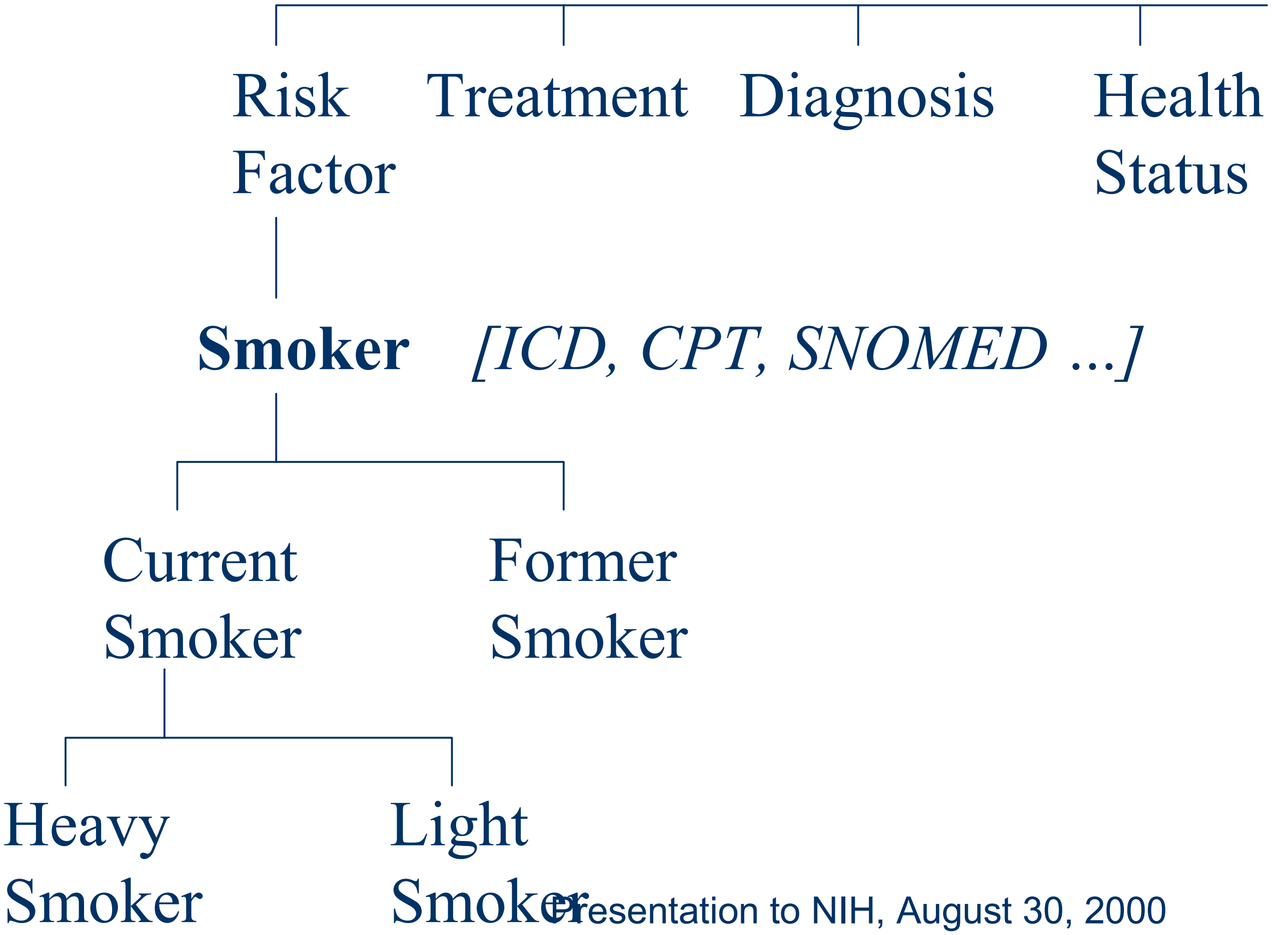
Dictionary

Research Database

Data Dictionary

- Concept-oriented: one medical concept may have many synonyms
- Integration of existing national and local vocabularies
- Based on firm knowledge representation
- Representation of source and conditions of collection

Dictionary Example



Risk Factor

Treatment

Diagnosis

Health Status

Smoker *[ICD, CPT, SNOMED …]*

Current Smoker

Former Smoker

Heavy Smoker

Light Smoker

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Protocol Management

- Recruitment for studies
- Enforcement of protocols
- Integration with administrative process

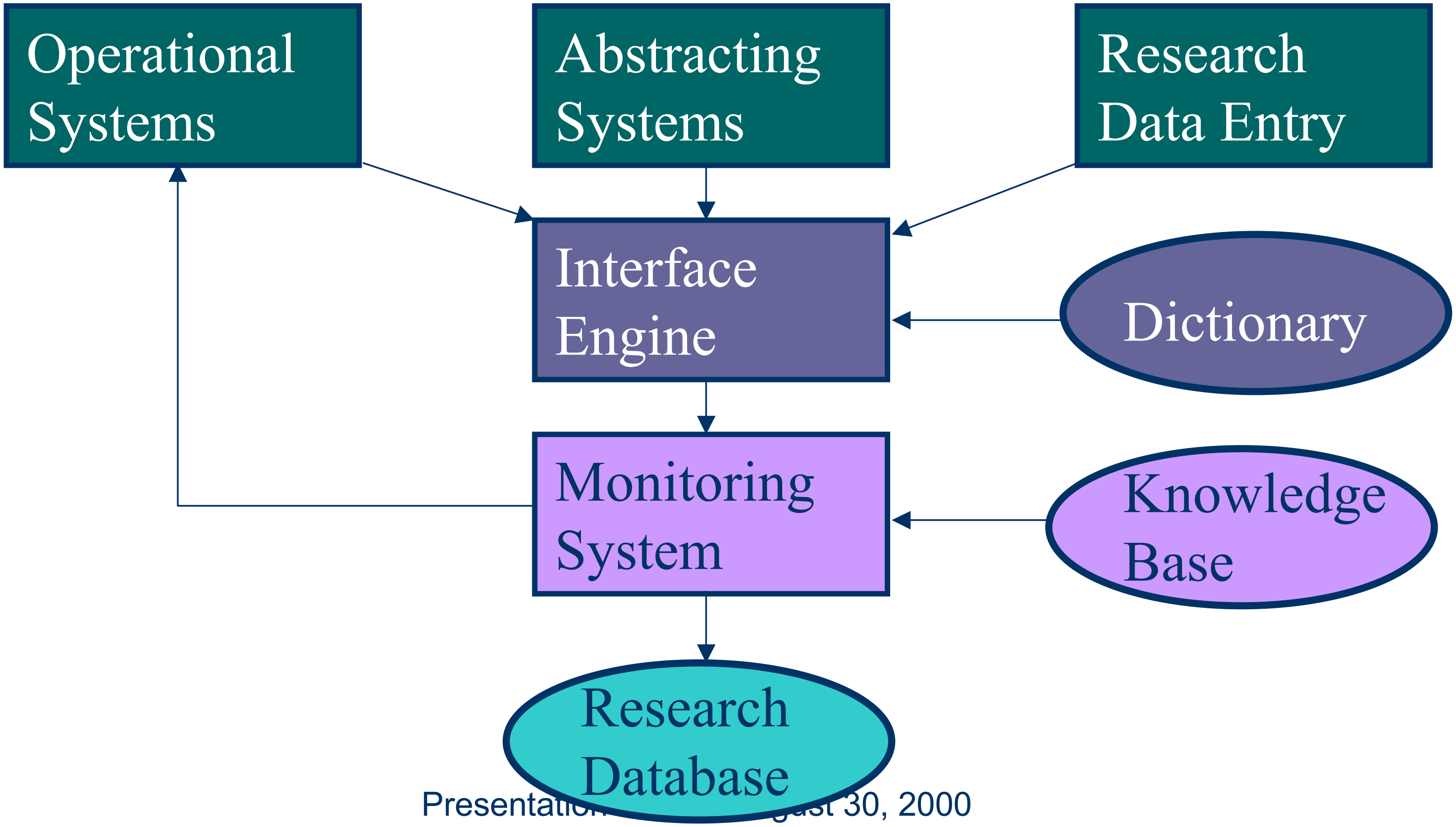
Protocol Management Systems

- ONCOCIN
  - Stanford University
- Oncology Center Information System
  - Johns Hopkins Oncology Center
- Proto-Direct
  - Dana Farber Cancer Center
- Protocol Data Management System
  - M.D. Anderson Cancer Center

Barriers to Automated Protocols

- Focus on single clinical domain
- Lack of integration with clinical systems
- Lack of integration with administrative systems
- Poor scalability for large populations and multiple protocols

Automated Clinical Monitoring



Automated Monitoring in Research

- Improve quality of research data
- Identify potential subjects for studies
- Notify researchers about events of interest

Automated Clinical Monitoring

- Knowledge about protocols is organized into a collection of “modules”
- Each clinical event is examined by monitor
- Modules relevant to an event are activated
- Modules generate alerts, warnings, reminders, notifications
- Messages are routed to appropriate personnel

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Research Data Collection

- Not in clinical record
- Have biases
- Not collected by appropriate person
- Time-consuming to collect
- Often in narrative form

Advanced Data Collection Techniques

- Protocol-directed data collection
- Structured data entry forms
- Speech recognition
- Natural language processing
- Standardization using dictionary
- Distribution of collection workload

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

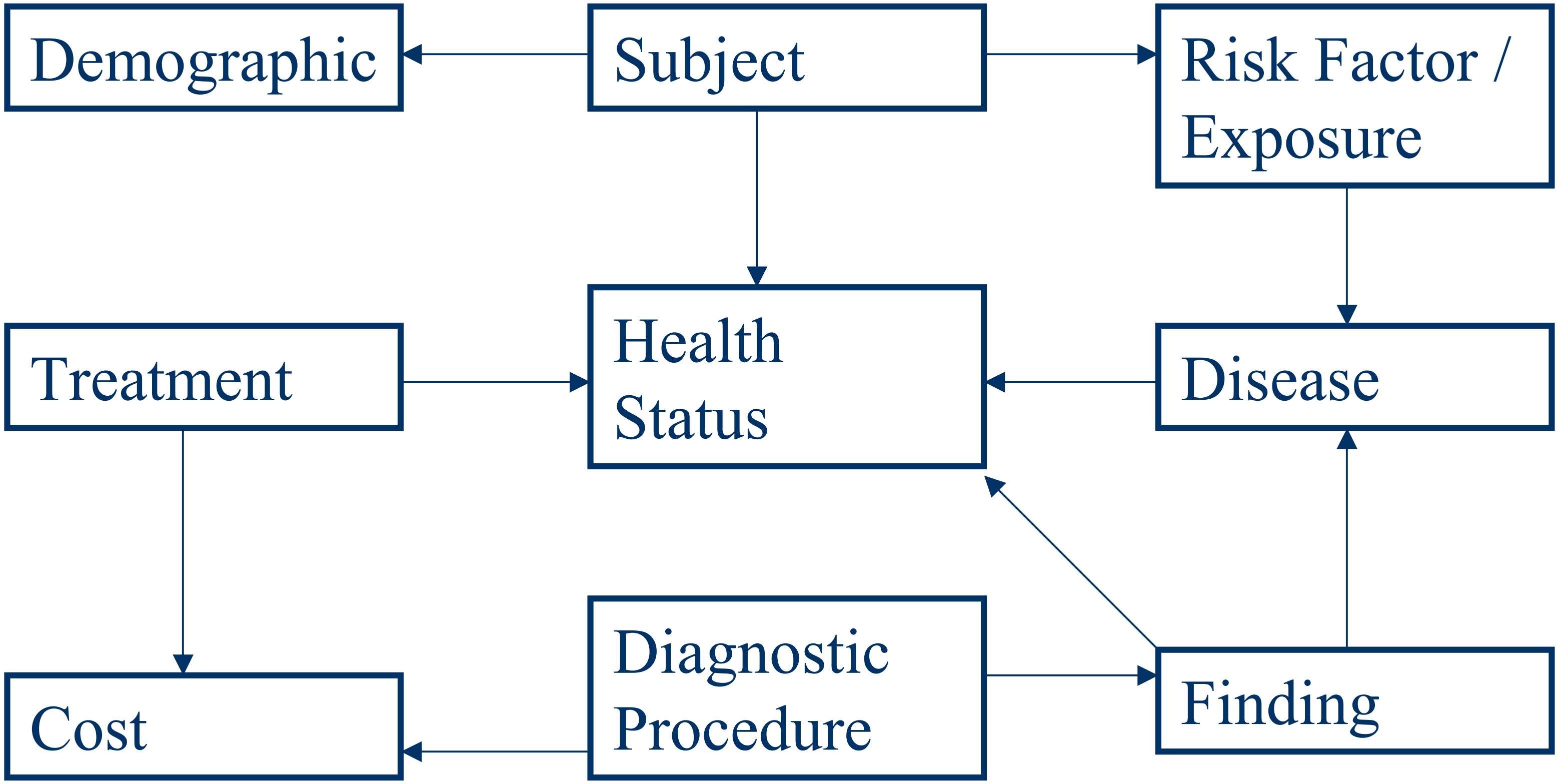
Design Issues

- Variety of data types
- Multiple granularities of data
- Flexible addition of new elements
- Efficiency for analytical processing (research queries)
- Scalability

Separate Research Database

- Different conceptual design
- Reduce impact on clinical systems
- Provide efficient research response
- Integrate with non-clinical data
- Protect research data

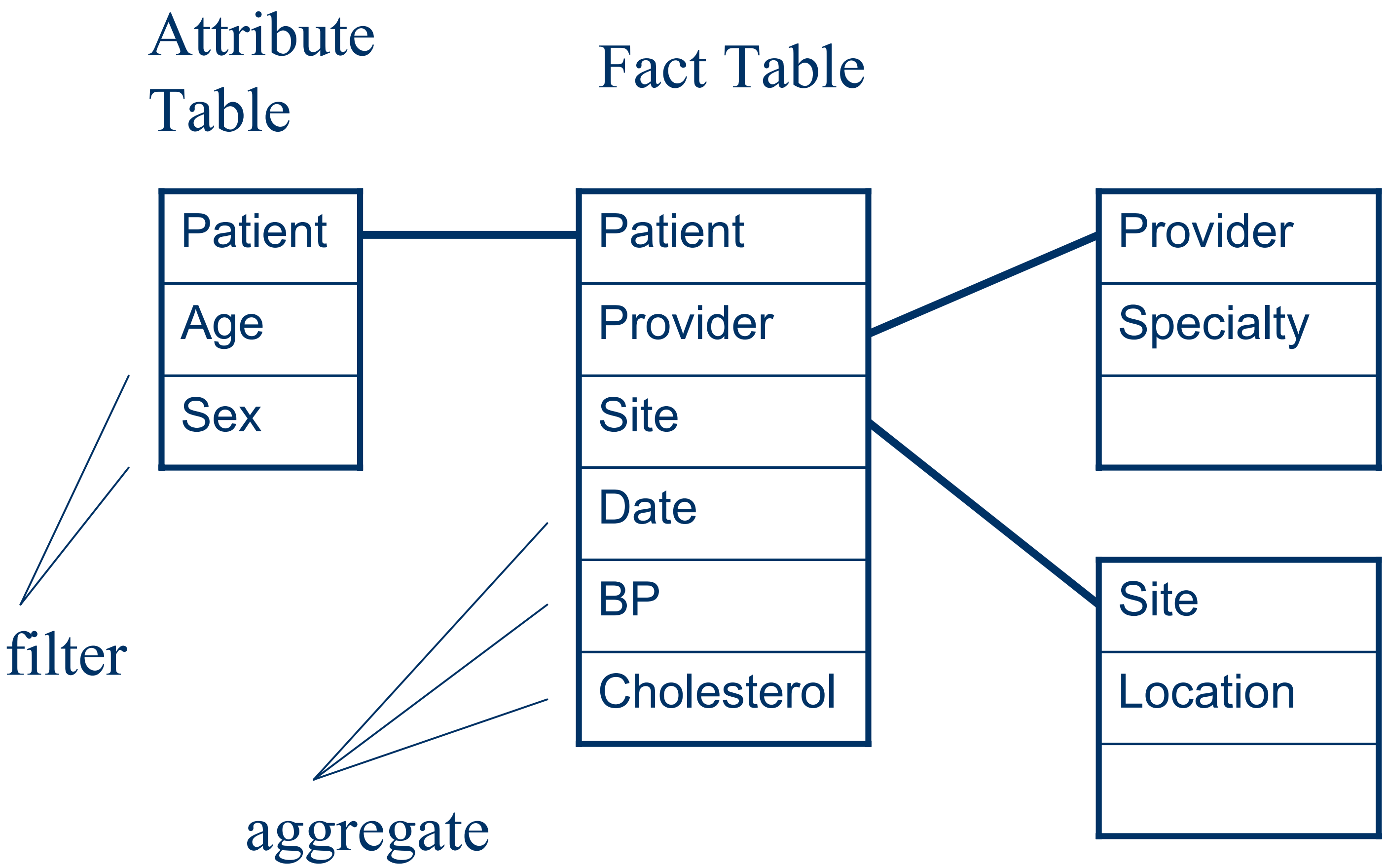
Reseach Database Schema



Extended Data Types

- Images
- Specimens
- Genetic Sequences
- Gene expression data

Database for Analytic Processing



Attribute Table

Fact Table

Patient
Age
Sex

Patient
Provider
Site
Date
BP
Cholesterol

Provider
Specialty

Site
Location

filter

aggregate

Individual Views

- Single table
- Relevant variables
- Appropriate granularity

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Analysis Tools

- Preformed reports save time
- Analytic database design is beneficial
- Query languages are complex
- Dictionary must be searchable
- Training is necessary
- Informatics support is necessary

Knowledge Discovery

- Data analysis require intense expert effort
- Potential of large data sets largely unknown
  - Unsupervised learning: no training set
- Hypotheses can be refined
  - Supervised learning: training set

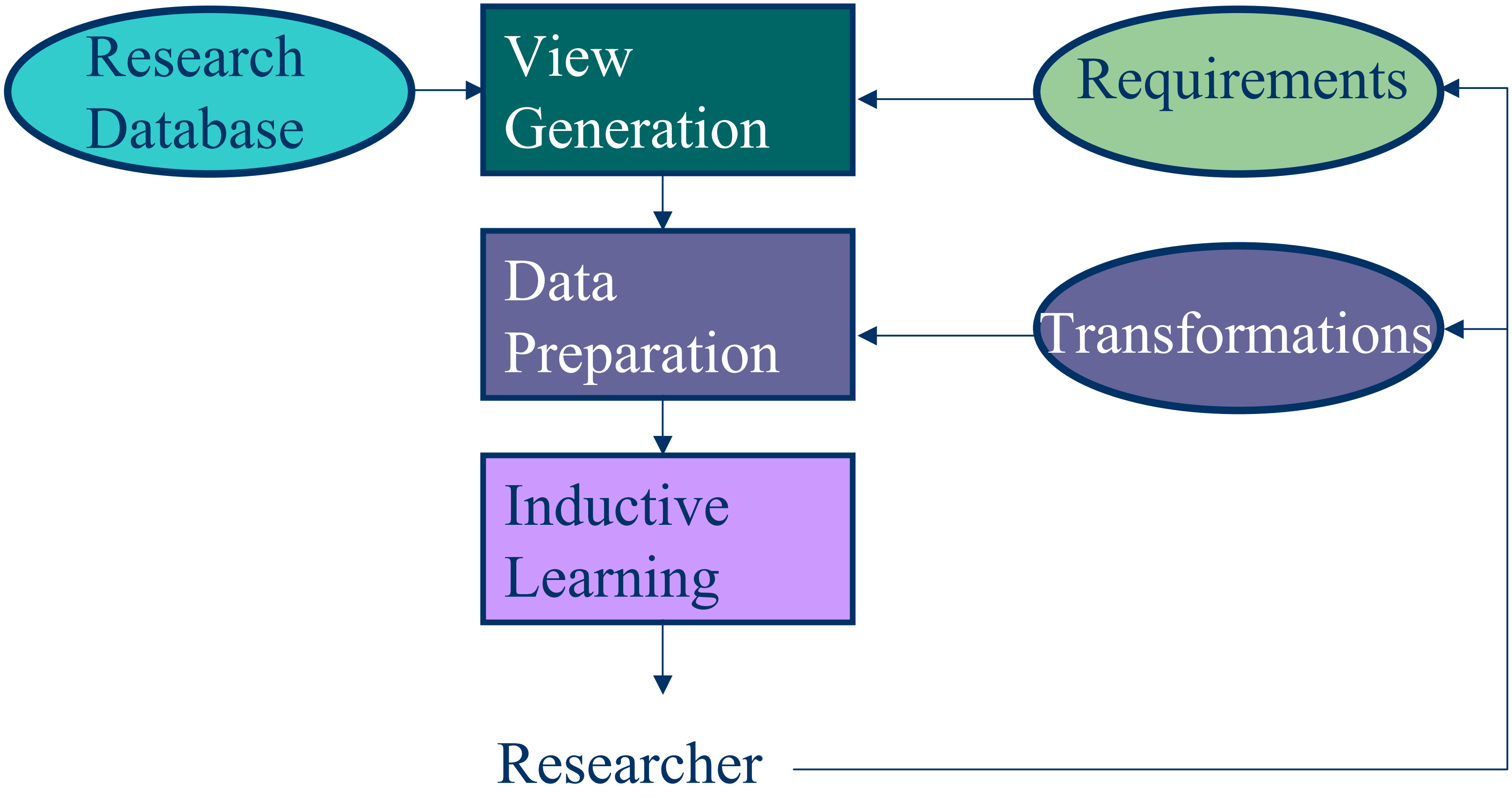
Discovery Methods

- Single table of variables
- Single outcome variable
- Data preparation – variable selection, discretizing, aggregation, imputation of null values
- Analysis with machine learning software

Machine Learning

- Technical Research (UC Irvine)
  - Small; artificial
- Clinical Research
  - Small; manually abstracted
- Administrative (MedisGroups, APACHE)
  - Large; manually abstracted
- Patient Care
  - Large; routinely collected

Data Analysis Architecture



Presentation to NIH, August 30, 2000

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Patient Privacy

- Health Insurance Portability and Accountability Act
- Department of Health and Human Services
- Protected Health Information
- Rules and penalties for disclosure

American Association of Medical Colleges

- Research data are not used to make decisions about individuals and therefore are not protected health information
- Research should continue to be regulated under Common Rule
- Research data should not be disclosed to patient or used in patient care
- Access of subject to clinical trial data should be determined by informed consent process
- Absolute right of access would prevent blinding and randomization
- Relevant clinical trial data related to an individual should be entered in patient record

Removal of Identifiers

- Eliminates possibility of benefit to patient
- Complicates maintenance of database
- Prevents auditing for fraud
- Distorts the data: subsampling, aggregation, noise introduction
- Fails to conceal, given sufficient facts

Research Database Design

- Create separate database for research
- Employ research identifier for each individual
- Maintain linkage to medical record in separate, secure database
- Use medical record to contact individuals for follow-up
- Retain specimen and procedure identifiers for linkage and access to specimens and images
- Scrub names, addresses, etc. from remaining fields

Protocol-based Protection

- Control access rather than content
- Approve each study through Institutional Review Board
- Require informed consent or waiver of consent when data are identifiable
- Support approval process through information systems
- Provide database “views” for each protocol

Protocol Views

- Access only to approved database columns
- View usable only by staff members conducting protocol
- Controlled by user identifier and password
- View active only for approved time period

Outline

- Introduction
- Architecture
- Variables
- Protocols
- Data Collection
- Database Design
- Analysis Tools
- Database Policy
- Organizational factors

Cost

- Interfaces and Dictionary - $2 million
- Data Management Professionals
  - 4 person hours / data element
  - $50 / hour
- Small Database (25 tables, 625 elements) - $3 million
- Medium Database (250 tables, 6250 elements) - $4 million
- Large Database(1000 tables, 25,000 elements) - $8 million

Diffusion of Innovation

User characteristics determine diffusion of innovation:

- Awareness of Resource
- Decision to Use
- Actual Use
- Continued Use

Success Factors

- Expend – provide sufficient resources
- Educate – advertise resources and support proficiency in computer applications, institutional coding and recording practices
- Enhance – provide data not available elsewhere
- Evolve – foster a culture of evaluation using shared data and methods of measurement

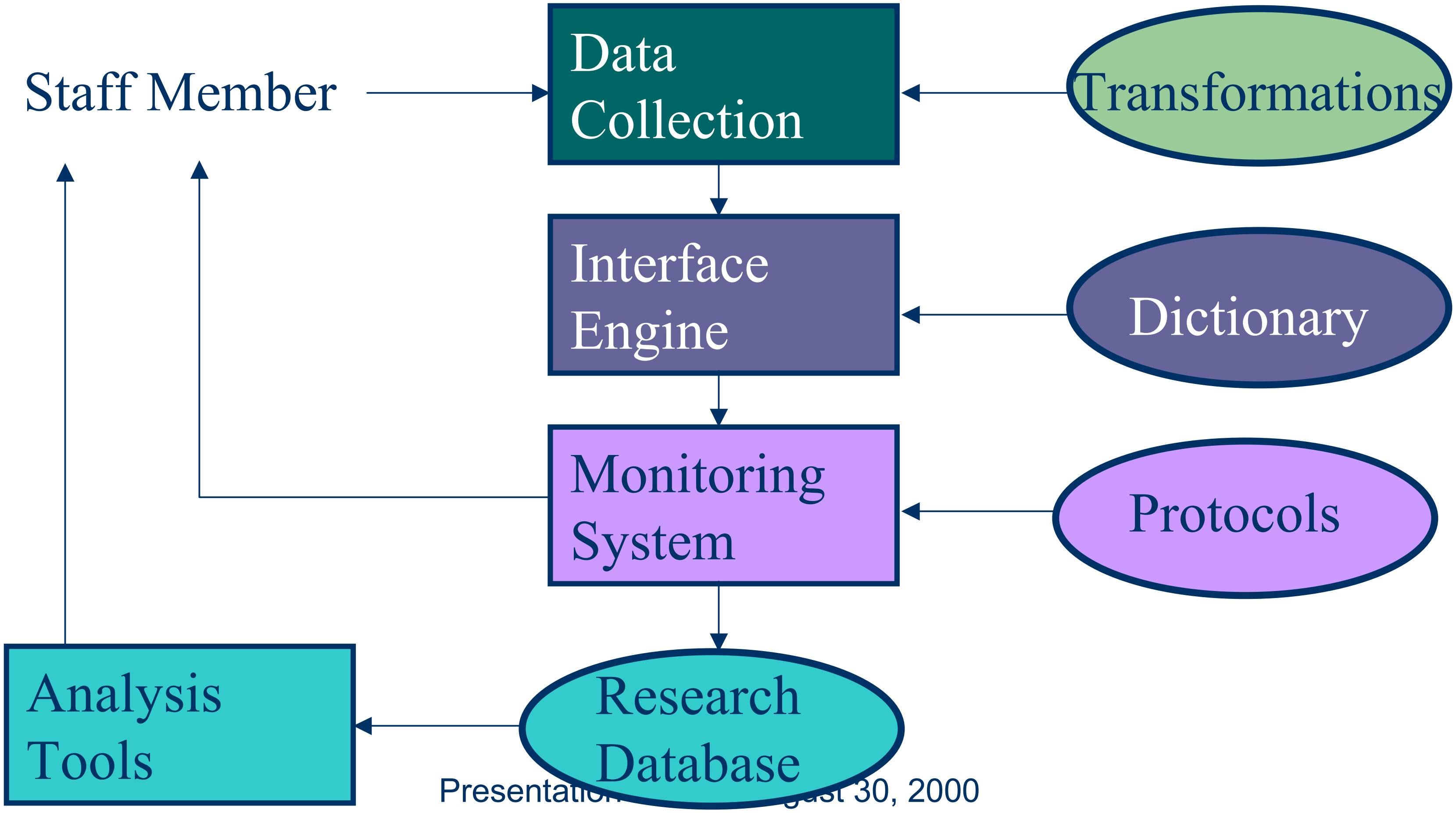
Conclusion

- Databases naturally proliferate
- Few institutions grasp advantage of pooling data
- Patient care and research are very different tasks
- Few successful systems combine both
- Current vendor products unlikely to scale

Design Requirements

- Open architecture – data exchange standards
- Dictionary – standardize variables and reduce biases
- Automating Monitoring – recruit subjects, implement protocols and direct data collection
- Research Database – independent research identifiers, efficiency for analysis
- Database Views – customized data access, study approval mechanism
- Organizational change – promote and support data sharing and analysis

Research Database Architecture


Staff Member
Data Collection
Transformations
Interface Engine
Dictionary
Monitoring System
Protocols
Analysis Tools
Research Database